Please cite this article as:

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35,* 2463-2476. doi:10.1016/j.ridd.2014.06.017

# Efficacy of Behavioral Interventions for Reducing Problem Behavior in Persons with Autism: An Updated Quantitative Synthesis of Single-Subject Research

Mieke Heyvaert[1, 2], Lore Saenen[1], Jonathan M. Campbell[3], Bea Maes[1], & Patrick Onghena[1]

[1] Faculty of Psychology and Educational Sciences - KU Leuven

[2] Postdoctoral Fellow of the Research Foundation - Flanders (Belgium)

[3] Department of Educational, School, and Counseling Psychology - University of Kentucky

Correspondence concerning this article can be addressed to Dr. Mieke Heyvaert, Methodology of Educational Sciences Research Group, Andreas Vesaliusstraat 2 - Box 3762, B-3000 Leuven, Belgium. Phone +32 16 326265. Fax +32 16 326200. E-mail Mieke.Heyvaert@ppw.kuleuven.be

## Abstract

Problem or challenging behaviors are highly prevalent among persons with autism and bring along major risks for the individual with autism and his/her family. In order to reduce the problem behavior, several behavioral interventions are used. We conducted a quantitative synthesis of single-subject studies to examine the efficacy of behavioral interventions for reducing problem behavior in persons with autism. Two hundred and thirteen studies representing 358 persons with autism met the inclusion criteria and were included in the statistical analyses. Overall, we found that behavioral interventions were on average effective in reducing problem behavior in individuals with autism, but some interventions were significantly more effective than others. The results further showed that the use of positive (nonaversive) behavioral interventions was increasing over time. The behavioral interventions were on average equally effective regardless of the type of problem behavior that was targeted. Interventions preceded by a functional analysis reduced problem behavior significantly more than interventions not preceded by a functional analysis. Finally, treatment and experimental characteristics, but not participant characteristics, were statistically significant moderators of the behavioral treatment effectiveness.



*Keywords:* meta-analysis, single-case, review, challenging behavior, problem behavior, intellectual disability

# Efficacy of Behavioral Interventions for Reducing Problem Behavior in Persons with Autism: An Updated Quantitative Synthesis of Single-Subject Research

## 1. Introduction

Problem behaviors such as aggressive, stereotyped, and self-injurious behavior are highly prevalent among persons with autism (e.g., Matson & LoVullo, 2008; Murphy, Healy, & Leader, 2009). The problem behaviors bring along major risks for the individual with autism and his/her family with regard to their physical, emotional, and social well-being, and can accordingly reduce their quality of life (e.g., Walsh, Mulder, & Tudor, 2013). In order to reduce problem behavior in persons with autism, several (cognitive-)behavioral interventions are used, such as differential reinforcement of other behavior (DRO), differential reinforcement of incompatible behavior (DRI), differential reinforcement of alternative behavior (DRA), antecedent control, antecedent exercise, noncontingent reinforcement, social stories, picture exchange communication system (PECS) interventions, and mindfulness-based interventions.

Many studies published in the domain of behavioral intervention research for reducing problem behavior among persons with autism are single-subject studies. In order to synthesize the results of these studies and to study which variables are moderating the effectiveness of the behavioral interventions, meta-level research is needed. Accordingly, Campbell (2003) conducted a quantitative synthesis of single-subject studies published between 1966 and 1998 on the efficacy of behavioral interventions for reducing problem behavior in persons with autism. In the meantime many more studies were published in this domain (cf. Matson & LoVullo, 2009), and an update of the meta-analysis of Campbell (2003) was warranted. The present article provides a double update of this meta-analysis: one from a methodological perspective and one from a temporal perspective.

First, we applied a methodological update. Campbell (2003) calculated three single-subject nonparametric statistics for estimating the effects of the behavioral treatments: the percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), the percentage of zero data (PZD; Scotti, Evans, Meyer, & Walker, 1991), and the mean baseline reduction (MBLR; Kahng, Iwata, & Lewin, 2002). However, in the meantime new single-subject nonparametric statistics have been developed that avoid some of the drawbacks of the earlier developed statistics (e.g., see Heyvaert, Wendt, Van den Noortgate, & Onghena, in press, Parker & Brossart, 2003, and Parker, Vannest, & Davis, 2011, for overviews). Therefore we included the percentage of data points exceeding the median of baseline phase (PEM; Ma, 2006) and the percentage of all nonoverlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007) in our update. In contrast with some other recently developed nonparametric statistics, PEM and PAND have comparable ease of use to PND, PZD, and MBLR (cf. Parker et al., 2011): All five effect sizes can easily be calculated by practitioners.

PND is the most frequently used effect size index across single-subject syntheses in the field of disability research (cf. Maggin, O'Keeffe, & Johnson, 2011). In order to meet PND's main drawback (i.e., the deficient performance in the presence of data outliers in the baseline phase) Ma (2006) developed PEM as an alternative effect size for summarizing results of single-subject studies: Whereas PND takes into account the one most extreme value of the baseline phase, and refers to the percentage of data points in the treatment phase that exceeds this most extreme value, PEM takes into account the median value of the baseline phase. PEM leans very close to PND in its calculation and interpretation. PAND was also developed to meet the drawbacks of PND, but conceptually deviates a bit more from it. The main advantages of PAND over PND are: (1) PAND uses all data from the baseline and intervention phases, avoiding the criticism leveled at PND for overemphasis on one unreliable data point; and (2) PAND can be translated to Pearson's *Phi* and $Phi^2$, and because *Phi* and

*Phi*$^2$ have known sampling distributions, *p* values are available, statistical power can be estimated, and confidence intervals can be included to indicate effect size reliability (Parker et al., 2007). Accordingly, for the present study we calculated PND, PZD, MBLR, PEM, and PAND for estimating the effects of behavioral interventions for reducing problem behavior in persons with autism. We compared these five nonparametric statistics with one another and examined to what extent they agreed in the analysis of the same data set. Answers to such questions are needed for scientist-practitioners to confidently use nonparametric statistics in the analysis of single-subject data (Parker & Brossart, 2003).

Second, for the temporal update we included single-subject studies published between 1999 and 2012 in our meta-analysis. Analogous to what Campbell (2003) did for the studies published between 1966 and 1998, we summarized single-subject studies published between 1999 and 2012: We studied the overall efficacy of behavioral interventions in reducing problem behavior in individuals with autism, examined whether some behavioral interventions were more effective than others, and investigated which variables, if any, moderated the overall efficacy of the behavioral interventions. Furthermore, we compared the two sets of studies and examined whether there were differences in the use of behavioral interventions and their effectiveness in reducing problem behavior in individuals with autism.

Accordingly, the research questions addressed in the current review were: (1) What is the overall efficacy of behavioral interventions in reducing problem behavior in individuals with autism; (2) Are some behavioral interventions more effective than others in reducing certain types of problem behavior in individuals with autism; (3) Do participant, treatment, or experimental variables influence the overall efficacy of behavioral interventions; (4) Are there any differences between the three older effect sizes (i.e., PND, PZD, and MBLR; Campbell, 2003) and the two more recently developed effect sizes (i.e., PEM and PAND) regarding treatment efficacy and moderating variables; and (5) Are there any differences between the

single-subject studies published between 1966 and 1998 (Campbell, 2003) and the studies published between 1999 and 2012 regarding the use of behavioral interventions and their effectiveness in reducing problem behavior in individuals with autism?

## 2. Method

### 2.1. Inclusion and exclusion criteria

We aimed at reviewing single-subject studies on behavioral interventions for reducing problem behavior in people with autism. The inclusion criteria were defined in the same way as Campbell (2003) did. First, the review included studies about participants diagnosed with autistic disorder. An article was included if at least one participant was diagnosed with autism. When articles included multiple individuals, only those participants diagnosed with autism were included in the review. Individuals described as 'autistic-like' or engaging in 'autistic-like behavior' were excluded. Second, studies were included if the behavioral treatment targeted reduction of self-injurious, stereotyped, or disruptive behavior, aggression, or property destruction. The third criterion concerned the study design: Only single-subject experiments were included that (a) described for each participant raw data points representing the level of problem behavior under baseline and treatment conditions, by intentional manipulation of the independent variable; (b) with the raw data points (not mean scores) for each participant separately reported in a table or clearly pictured in a graph; and (c) with baseline and treatment conditions containing at least two data points for each participant (cf. Campbell, 2003). Accordingly, group comparison studies that did not report raw data at the individual participant level were excluded. The fourth criterion concerned the time period. In the review conducted by Campbell (2003), studies published between 1966 and 1998 were included. Because the present review was an update of the review of Campbell (2003), single-subject studies published between 1999 and 2012 were included. We started the systematic

search for studies in 2013. Fifth, the articles had to be written in English in order to be understood by the research team.

**2.2. Systematic search process for original studies**

Studies were retrieved by systematically searching several electronic databases, relevant journals, bibliographies of relevant articles, and citation indexes. First, we searched seven relevant electronic databases: Academic Search Elite (ASE), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Embase, Education Resources Information Center (ERIC), Medline, PsycINFO, and Web of Science. We used the following search string: (autism OR autistic) AND (aggression OR behavio* problems OR challenging behavio* OR destructive behavio* OR disruptive behavio* OR problem behavio* OR property destruction OR repetitive behavio* OR self-harm OR self-injurious behavio* OR self-injury OR self-stimulation OR SIB OR stereotyped behavio* OR stereotypy). Second, we conducted a hand search of 26 relevant journals: *American Journal on Intellectual and Developmental Disabilities (American Journal on Mental Retardation); Behavior Modification; Behavior Research and Therapy; Behavior Therapy; Behavioral Interventions; British Journal of Learning Disabilities; Child and Family Behavior Therapy; Clinical Case Studies; Focus on Autism & Other Developmental Disabilities; Intellectual and Developmental Disabilities (Mental Retardation); International Journal of Rehabilitation Research; Journal of Abnormal Child Psychology; Journal of Applied Behavior Analysis; Journal of Applied Research in Intellectual Disabilities; Journal of Autism and Developmental Disorders; Journal of Behavior Therapy and Experimental Psychiatry; Journal of Clinical Child & Adolescent Psychology; Journal of Consulting and Clinical Psychology; Journal of Developmental and Physical Disabilities; Journal of Experimental Child Psychology; Journal of Intellectual and Developmental Disability; Journal of*

*Intellectual Disabilities; Journal of Intellectual Disability Research; Journal of Positive Behavior Interventions; Research in Autism Spectrum Disorders;* and *Research in Developmental Disabilities.* Third, we examined the bibliographies of all articles that were identified as relevant in the first and second search step. Fourth, we searched for more recently published studies referring to the papers identified as relevant in the three previous search steps, by consulting three citation indexes: the Arts & Humanities Citation Index, the Science Citation Index Expanded, and the Social Sciences Citation Index - all three accessed through Web of Science. Our systematic search process identified 213 studies describing data of 358 participants that met all inclusion criteria. The complete list of the articles included in this review can be requested from the first author.

### 2.3. Estimating effects of behavioral treatments

For the present study we calculated five single-subject nonparametric statistics for estimating the effects of the behavioral treatments: PND, PZD, MBLR, PEM, and PAND (cf. **Introduction**). PND is calculated by dividing the number of treatment data points that exceeds the highest baseline data point in the expected direction by the total number of data points in the treatment phase (Scruggs et al., 1987). For studies on reducing problem behavior, PND is the percentage of data points in the treatment phase that is lower than the lowest baseline point. A PND score higher than 90% indicates a highly effective treatment, a score between 90% and 70% an effective treatment, a score between 70% and 50% a questionable treatment, and a score lower than 50% indicates an ineffective treatment.

PZD is calculated by locating the first data point in the treatment phase that reaches zero and calculating the percentage of data points recorded in the treatment phase, including the first zero, that remains at zero (Scotti et al., 1991). A PZD score higher than 80% indicates a highly effective treatment, a score between 80% and 55% an effective treatment, a score

between 55% and 18% a questionable treatment, and a score lower than 18% indicates an ineffective treatment.

MBLR is calculated by subtracting the mean treatment value from the mean baseline value, and dividing by the mean baseline value (Kahng et al., 2002). An MBLR score of 100% reflects total elimination of the problem behavior, whereas a 0% score reflects no change from baseline. A negative MBLR score reflects an increase in the problem behavior during treatment.

PEM is calculated by dividing the number of treatment data points that exceeds the median baseline data point in the expected direction by the total number of data points in the treatment phase (Ma, 2006). For studies on reducing problem behavior, PEM is the percentage of data points in the treatment phase that is lower than the median baseline point. The interpretation of PEM is parallel to PND: A PEM score higher than 90% indicates a highly effective treatment, a score between 90% and 70% an effective treatment, a score between 70% and 50% a questionable treatment, and a score lower than 50% indicates an ineffective treatment.

Finally, PAND refers to the percentage of data points that does not overlap between the baseline and treatment phases, and is calculated as follows: (1) Identify the number of overlapping data points (i.e., the minimum number of data points that would have to be transferred across phases for complete data separation), (2) compute the percentage overlap by dividing the number of overlapping points by the total number of data points, and (3) subtract this percentage from 100% (Parker et al., 2007). PAND is scaled from 50% to 100%, where 50% is chance level (cf. Parker et al., 2011).

**Figure 1** presents an empirical example, illustrating how to calculate these five effect sizes.

**INSERT FIGURE 1 ABOUT HERE**

Information on the effectiveness of the behavioral treatments was gathered from the graphs presented in the retrieved single-subject articles. The raw data (i.e., XY-coordinates of all data points in the graphs) were extracted using UnGraph Version 5 (Biosoft, 1997-2014), a software program proven to show highly valid and reliable data extraction results (Shadish et al., 2009). The five effect sizes were calculated for each participant and each study that met the inclusion criteria (cf. **2.1**).

With regard to the weighing of the effect sizes, we followed the procedure described by Campbell (2003): When more than one problem behavior was targeted for a participant, the average effect sizes for the participant were calculated by weighting each behavior according to the number of data points reporting on the behavior. Within each article, effect sizes were weighted according to the number of data points per participant and then averaged for all participants to yield five effect sizes per article. All effect sizes were calculated by comparing the first baseline phase to the final treatment phase.

### 2.4. Predictor variables coded and reliability

For each participant we extracted data on three groups of characteristics: participant characteristics, characteristics of the behavioral treatment used, and experimental characteristics. First, we coded five participant characteristics: age (in years), gender, criteria used for diagnosing autism, intellectual disability level, and level of verbal communication ability. Second, six treatment characteristics were coded: type of intervention used, type of problem behavior that was targeted, parental involvement in the intervention, functional analysis, availability of follow up data, and whether efforts were made to generalize the behavior change. Third, we coded five experimental characteristics: experimental design used,

number of coded baseline data points, number of coded treatment data points, publication year, and whether or not inter-rater reliability data were reported. More information on the coded participant, treatment, and experimental variables can be found in **Tables 1**, **2**, and **3** respectively.

Both the first and second author coded 14 variables for all included participants ($N$ = 358). The variables 'number of coded baseline data points' and 'number of coded treatment data points' were generated when extracting the raw data points from the graphs (cf. **2.3**). As part of the coding training, the two authors together coded all 14 variables for the first, second, and third included participant. The first and second author independently coded the 14 variables for the remaining 355 participants. Inter-rater agreement was calculated for the 14 x 355 cells by dividing the number of agreements by the number of agreements plus disagreements. The inter-rater agreement was 99.82%. Disagreements were afterwards resolved by discussion between the first and second author, and the corrected codes were used for the analyses.

### 2.5. Statistical analyses

Following Campbell (2003), the analyses of the main intervention effects were conducted at the participant ($N$ = 358) as well as at the study level ($N$ = 213). We studied five effect sizes: PND, PZD, MBLR, PEM, and PAND (cf. **2.3**). All statistical analyses related to the predictor variables were conducted at the participant level. In order to answer the three first research questions, three groups of statistical analyses were conducted. First, the overall efficacy of the behavioral interventions was examined by calculating the mean effect sizes at the participant and study level. In addition, we constructed 95% confidence intervals around the mean effect sizes. Second, five two-way analyses of variance (ANOVAs) were used to test for the main effects of treatment type and target behavior as well as their interaction for

each effect size. Third, we conducted a series of hierarchical multiple regression analyses. We conducted these analyses for the three groups of characteristics: participant characteristics, characteristics of the behavioral treatment used, and experimental characteristics. For analyzing all three groups of characteristics, we used the backward stepwise method, with the participant/treatment/experimental characteristics entered first, and behavior category and treatment type entered afterwards. The analyses were conducted separately for the five effect sizes. Due to the presence of multiple contrasts, Bonferroni corrections were used in order to control for familywise error rates. For the five two-way ANOVAs as well as for the hierarchical multiple regression analyses, results associated with *p* values below .003 (.05/15) were interpreted as statistically significant. We used the SPSS software (Version 22; SPSS Inc., 2013-2014) to conduct all statistical analyses.

## 3. Results

Two hundred and thirteen studies representing 358 persons with autism met the inclusion criteria and were included in the statistical analyses. Detailed information about the participants, treatments, and experimental studies is presented in **Tables 1**, **2**, and **3** respectively.

**INSERT TABLES 1, 2, 3 ABOUT HERE**

To answer the first research question we examined the overall efficacy of the behavioral interventions. Across all participants the averages were 74.9%, 44.7%, 70.2%, 91.4% and 91.9% for PND, PZD, MBLR, PEM, and PAND respectively. At the study level the averages were 75.9%, 47.3%, 74.2%, 93.0%, and 92.3% respectively. The means and standard deviations and the 95% confidence intervals for the five effect sizes at the participant

as well as at the study level are presented in **Table 4**. Following the interpretation guidelines for the effect sizes (cf. **2.3**), we can conclude for PND, MBLR, PEM, and PAND that the behavioral treatments were on average effective in reducing problem behavior at both participant and study level. However, for PZD the mean averages at participant and study level were below 55%, indicating questionable treatment effects (cf. **2.3**).

**INSERT TABLE 4 ABOUT HERE**

The behaviors most often targeted by the behavioral interventions were stereotyped behavior only (27.9%), disruptive behavior only (22.9%), a combination of internalized and externalized problem behaviors (21.8%), a combination of externalized problem behaviors (10.3%), self-injurious behavior only (9.2%), and aggression only (7.3%). The behavioral interventions most often used were combinations of positive interventions (31.8%), combinations of aversive and positive interventions (20.7%), antecedent control (17.0%), social stories (8.1%), and DRA (6.4%) interventions. Treatment type and target behavior were not independent of one another, $\chi^2(84, N = 358) = 264.14, p < .001$. Stereotyped behaviors and combined internalized and externalized problem behaviors tended to be treated mostly with a combination of positive and aversive behavioral techniques (respectively: $n = 31$; $n = 22$) as well as with a combination of positive behavioral techniques (respectively: $n = 19$; $n = 42$). In addition, stereotyped behaviors tended to be treated mostly with antecedent control interventions ($n = 31$). Furthermore, disruptive problem behavior tended to be treated mostly with social stories interventions ($n = 25$).

To answer the second research question five two-way ANOVAs were used to test for the main effects of treatment type and target behavior as well as their interaction for each of

the five effect sizes (cf. **Table 5**). Due to the large number of contrasts involved in the ANOVAs, Bonferroni correction was used to determine whether significant main effects and interactions were present in these analyses. Therefore, results associated with *p* values below .003 (.05/15) were interpreted as statistically significant. There was a statistically significant main effect for treatment type for three effect sizes: for PND, $F(12, 307) = 3.95$, $p < .001$; for PEM, $F(12, 307) = 3.32$, $p < .001$; and for PAND, $F(12, 307) = 3.64$, $p < .001$. After applying the Bonferroni correction there was no statistically significant main effect for treatment type for PZD, $F(12, 307) = 2.03$, $p = .021$, nor for MBLR, $F(12, 307) = 2.05$, $p = .020$. No main effect was found for target behavior for any of the five effect sizes: for PND, $F(7, 307) = 1.28$, $p = .262$; for PZD, $F(7, 307) = 1.54$, $p = .152$; for MBLR, $F(7, 307) = 0.91$, $p = .503$; for PEM, $F(7, 307) = 1.18$, $p = .313$; and for PAND, $F(7, 307) = 0.80$, $p = .592$. Prior to the Bonferroni correction, there was only a statistically significant interaction between treatment type and target behavior for PND, $F(31, 307) = 1.50$, $p = .047$. However, after applying the Bonferroni correction, the interaction between treatment type and target behavior for PND was no longer statistically significant. For the four other effect sizes, there was clearly no statistically significant interaction between treatment type and target behavior (see **Table 5**). Levene's test of equality of error variance was statistically significant for all five effect sizes: for PND, $F(50, 307) = 2.99$, $p < .001$; for PZD, $F(50, 307) = 2.07$, $p < .001$; for MBLR, $F(50, 307) = 2.51$, $p < .001$; for PEM, $F(50, 307) = 4.86$, $p < .001$; and for PAND, $F(50, 307) = 3.88$, $p < .001$. Because the assumption of homogeneity of variances had been violated, a series of independent samples Kruskal-Wallis tests was used to test for main effects of treatment type and problem behavior type for the five effect sizes. For treatment type the null hypothesis was rejected for all effect sizes ($p < .001$ for all five). For type of problem behavior the null hypothesis was rejected for PZD only ($p < .001$).

**INSERT TABLE 5 ABOUT HERE**

Afterwards, we explored in more detail the statistically significant main effect for treatment type that was found for the effect sizes PND, PEM, and PAND: We used pairwise post hoc contrasts and applied the Tukey-Kramer corrections. Because there was only one participant treated with an escape only intervention (cf. **Table 2**), this treatment type could not be included in the analyses. For PND we found eight statistically significant contrasts: One of the contrasts concerned antecedent control only interventions and the seven other contrasts concerned PECS only interventions. Positive combination interventions (cf. **Table 2**) were statistically significantly better in reducing problem behavior than antecedent control only interventions. Seven treatment types were statistically significantly better in reducing problem behavior than PECS only interventions: aversive and positive combinations, positive combinations, DRO only, antecedent control only, DRA only, noncontingent reinforcement only, and social stories only interventions. The results for the pairwise post hoc contrasts for the effect sizes PEM and PAND were identical: We found 13 statistically significant contrasts, two of them concerning antecedent control only interventions and the 11 other concerning PECS only interventions. Aversive and positive combinations as well as positive combination interventions were statistically significantly better in reducing problem behavior than antecedent control only interventions. Eleven treatment types were statistically significantly better in reducing problem behavior than PECS only interventions: aversive and positive combinations, positive combinations, punishment only, DRO only, antecedent control only, DRI only, DRA only, antecedent exercise only, noncontingent reinforcement only, social stories only, and mindfulness-based strategy only interventions.

For 71.8% of the participants a functional analysis was conducted prior to the behavioral intervention. Publication year and presence of functional analysis were not

independent, $\chi^2(13, N = 358) = 35.93, p = .001$. There was an increase in the use of pretreatment functional analysis over time. Presence of functional analysis and intervention type were also not independent, $\chi^2(12, N = 358) = 144.17, p < .001$. The interventions that were more likely to be preceded by a functional analysis were noncontingent reinforcement interventions (100%), antecedent exercise interventions (100%), DRA interventions (95.7%), combinations of positive interventions (89.5%), combinations of aversive and positive interventions (86.5%), and DRO interventions (71.4%). Consistent with the trends for functional analysis, publication year and intervention type were not independent, $\chi^2(156, N = 358) = 277.69, p < .001$, with an increase in the use of positive interventions over time.

Given the empirical evidence for the benefit of pretreatment functional analysis (e.g., Campbell, 2003; Didden, Korzilius, van Oorsouw, & Sturmey, 2006; Harvey, Boer, Meyer, & Evans, 2009; Scotti et al., 1991), the influence of functional analysis was tested via five one-way ANOVAs with the effect sizes entered as dependent variables and the presence of functional analysis as the independent variable. Due to the presence of multiple contrasts, Bonferroni correction was again used to determine statistical significance: Results associated with *p* values below .01 (.05/5) were interpreted as statistically significant. The ANOVAs revealed a statistically significant effect for PND scores, $F(1, 356) = 13.75, p < .001$; for MBLR scores, $F(1, 356) = 11.45, p = .001$; for PEM scores, $F(1, 356) = 22.19, p < .001$; and for PAND scores, $F(1, 356) = 15.82, p < .001$. No main effect was found for PZD scores, $F(1, 356) = 0.26, p = .610$. Examination of mean scores revealed that the scores for all five effect sizes were higher for studies including a pretreatment functional analysis ($M = 78.9\%$, $SD = 30.5\%$ for PND; $M = 45.3\%$, $SD = 36.6\%$ for PZD; $M = 74.6\%$, $SD = 35.2\%$ for MBLR; $M = 94.2\%$, $SD = 12.2\%$ for PEM; $M = 93.3\%$, $SD = 9.8\%$ for PAND), compared to studies not including a pretreatment functional analysis ($M = 64.6\%$, $SD = 38.5\%$ for PND; $M = 43.1\%$,

*SD* = 33.7% for PZD; *M* = 59.0%, *SD* = 47.9% for MBLR; *M* = 84.1%, *SD* = 28.2% for PEM; *M* = 88.2%, *SD* = 13.3% for PAND).

For respectively 37.2% and 20.7% of the participants attempts to generalize behavior change and follow-up data were reported. Although the presence of follow-up data was significantly related to the publication year, $\chi^2(13, N = 358) = 30.04$, $p = .005$, there was no statistically significant trend over time for attempts to generalize behavior change, $\chi^2(13, N = 358) = 14.66$, $p = .329$.

**INSERT TABLE 6 ABOUT HERE**

To answer the third research question we conducted hierarchical multiple regression analyses for the three groups of characteristics: participant characteristics, characteristics of the behavioral treatment used, and experimental characteristics. For all three analyses, we used the backward stepwise method, with the participant/treatment/experimental characteristics entered first, and behavior category and treatment type entered afterwards. The analysis started with all predictors included in the model. Next, it was tested whether any of these predictors – except behavior category and treatment type – could be removed without having a substantial effect on how well the model fits the data. The analyses were conducted separately for the five effect sizes. The results of the hierarchical multiple regression analyses are presented in **Table 6**.

In the first group of analyses, the participant characteristics age, gender, criteria used for diagnosing autism, intellectual disability level, and level of verbal communication ability were entered first, and behavior category and treatment type were entered afterwards. For PND and PAND there was a statistically significant *F Change* for the model including all seven predictors, respectively *F Change*(7, 345) = 2.22, $p = .032$; *F Change*(7, 345) = 2.10, *p*

= .043. However, due to the presence of multiple contrasts, Bonferroni correction was used; therefore, results associated with *p* values below .003 (.05/15) were interpreted as statistically significant. After applying the Bonferroni correction for PND and PAND there was no longer a statistically significant *F Change* for any model. For PZD there was no statistically significant *F Change* for any of the five models: The analyses suggested working with Model 5 including the predictors intellectual disability level, behavior category, and treatment type. Similarly, for MBLR and PEM there was no statistically significant *F Change* for any of the six models: The analyses for both effect sizes suggested working with Model 6 including behavior category and treatment type. Likewise, after applying the Bonferroni correction for PND and PAND there was no statistically significant *F Change* for any of the six models: The analyses suggested working with the model only including behavior category and treatment type.

In the second group of analyses, the treatment characteristics parental involvement, functional analysis, presence of follow up data, and efforts to generalize behavior change were entered first, and behavior category and treatment type were entered afterwards. Prior to applying the Bonferroni correction for all five effect sizes there was a statistically significant *F Change* for the model including all six predictors, $F\ Change(6, 351) = 6.69$, $p < .001$ for PND; $F\ Change(6, 351) = 2.60$, $p = .018$ for PZD; $F\ Change(6, 351) = 3.89$, $p = .001$ for MBLR; $F\ Change(6, 351) = 6.34$, $p < .001$ for PEM; $F\ Change(6, 351) = 5.52$, $p < .001$ for PAND. However, after applying the Bonferroni correction there was only a statistically significant *F Change* for the model including all six predictors for PND, MBLR, PEM, and PAND.

In the third group of analyses, the experimental characteristics design, number of coded baseline data points, number of coded treatment data points, publication year, and inter-rater reliability were entered first, and behavior category and treatment type were entered

afterwards. The results for the experimental characteristics were analogous to the findings for the treatment characteristics: Prior to applying the Bonferroni correction for all five effect sizes there was a statistically significant *F Change* for the model including all seven predictors, *F Change*(7, 350) = 12.90, $p < .001$ for PND; *F Change*(7, 350) = 2.15, $p = .038$ for PZD; *F Change*(7, 350) = 4.22, $p < .001$ for MBLR; *F Change*(7, 350) = 5.72, $p < .001$ for PEM; *F Change*(7, 350) = 10.22, $p < .001$ for PAND. However, after applying the Bonferroni correction there was only a statistically significant *F Change* for the model including all six predictors for PND, MBLR, PEM, and PAND. The PZD analyses suggested working with Model 4 including publication year, inter-rater reliability, behavior category, and treatment type.

The fourth research question concerned the comparison of the five calculated effect sizes: We wanted to examine for the single-subject studies published between 1999 and 2012 whether there were differences in the conclusions relating to treatment efficacy and moderating variables for the five effect sizes. The analyses for PND, MBLR, PEM, and PAND resulted in similar conclusions on treatment efficacy and moderator analyses (cf. supra), whereas PZD seemed to be the "odd man out". First of all, with regard to the overall effect sizes we found that according to PND, MBLR, PEM, and PAND the behavioral treatments were on average effective in reducing participants' problem behavior, whereas for PZD the mean averages at both participant and study level indicated questionable treatment effects. Second, PZD offered different conclusions than the four other effect sizes for the moderator analyses. Based on the hierarchical multiple regression analyses we found that, after applying Bonferroni corrections, treatment and experimental characteristics did significantly influence the overall efficacy of the behavioral interventions for PND, MBLR, PEM, and PAND, but not for PZD. Another example was given by the results for functional

analysis: We found a statistically significant relation between the overall intervention effect and the presence of functional analysis for PND, MBLR, PEM, and PAND, but not for PZD.

The final research question concerned the comparison of the single-subject studies published between 1966 and 1998 (Campbell, 2003) with the studies published between 1999 and 2012. First of all, we wanted to examine whether there were differences in the use of behavioral interventions for reducing problem behavior in individuals with autism between the two sets of studies. The behavioral interventions most often used for the studies published between 1966 and 1998 (Campbell, 2003) were combinations of aversive and positive interventions (23.9%), DRO (12.8%), punishment (11.1%), antecedent control (8.5%), positive combinations (6.8%), and overcorrection (6.8%) interventions. The interventions most often used for the studies published between 1999 and 2012 were combinations of positive interventions (31.8%), combinations of aversive and positive interventions (20.7%), antecedent control (17.0%), social stories (8.1%), and DRA (6.4%) interventions. Comparing the two sets, we observed an increase in the use of positive interventions over time, together with a decrease in the use of negative interventions. Looking at the problem behavior types most often targeted by behavioral interventions, we observed parallel results for both sets, with two exceptions: Self-injurious behavior only was more often targeted in the studies published between 1966 and 1998 (17.9%, versus 9.2% for the present data set), and disruptive behavior only was more often targeted in the studies published between 1999 and 2012 (22.9%, versus 7.7% the studies published between 1966 and 1998).

Second, we wanted to examine whether smaller or larger intervention effects were reported in the most recent set of studies, compared to the older set of studies. For both data sets it was found that, except for the PZD results, the behavioral treatments were on average effective in reducing the participants' problem behavior. The overall efficacy of the

behavioral interventions for the studies published between 1966 and 1998 was 84.4%, 42.9%, and 76.5% for PND, PZD, and MBLR respectively (Campbell, 2003). For the studies published between 1999 and 2012 study level averages were 75.9%, 47.3%, and 74.2% respectively. Thus, the average MBLR effect size was quite similar for both data sets, the average PND score was a bit higher for the oldest set of studies, and the average PZD score was a bit higher for the most recent set of studies.

## 4. Discussion

The present study aimed to answer five questions: (1) What is the overall efficacy of behavioral interventions in reducing problem behavior in individuals with autism; (2) Are some behavioral interventions more effective than others in reducing certain types of problem behavior in individuals with autism; (3) Do participant, treatment, or experimental variables influence the overall efficacy of behavioral interventions; (4) Are there any differences in the conclusions for the five calculated effect sizes regarding treatment efficacy and moderating variables; and (5) Are there any differences between the single-subject studies published between 1966 and 1998 (Campbell, 2003) and the studies published between 1999 and 2012 regarding the use of behavioral interventions and their effectiveness in reducing problem behavior in individuals with autism?

For the first question, we conclude for PND, MBLR, PEM, and PAND that the behavioral treatments were on average effective in reducing problem behavior at both the participant and study level. This conclusion corresponds to the findings of other meta-analyses published in this domain: Behavioral treatments are on average effective in reducing problem behavior in individuals with autism (e.g., Didden et al., 2006; Harvey et al., 2009; Heyvaert, Maes, & Onghena, 2010; Heyvaert, Maes, Van den Noortgate, Kuppens, &

Onghena, 2012; Heyvaert, Saenen, Maes, & Onghena, 2014; Scotti et al., 1991; Vanderkerken, Heyvaert, Maes, & Onghena, 2013). However, for PZD the mean averages at the participant and study level indicated questionable treatment effects. The difference between PZD and PND, MBLR, PEM, and PAND will be further discussed below.

With regard to the second research question, we found a statistically significant main effect for treatment type for three of the five effect sizes: PND, PEM, and PAND. We examined these results in further depth using pairwise post hoc contrasts and found eight statistically significant contrasts for PND and 13 for PEM and PAND. All the contrasts concerned antecedent control only interventions or PECS only interventions. Aversive and positive combinations as well as positive combination interventions were statistically significantly better in reducing problem behavior than antecedent control only interventions. Furthermore, almost all the treatment types included in the present study were statistically significantly better in reducing problem behavior than PECS only interventions. We note that the number of participants treated by PECS only interventions that was included in the analyses was very small (i.e., 5 participants; cf. **Table 2**). Accordingly, this intervention should not yet be written off as a stand-alone intervention for reducing problem behavior in persons with autism. Future empirical research on PECS interventions should be critically evaluated. Our findings contrast with the results of Campbell (2003) for the single-subject studies published between 1966 and 1998, for which no main effect for treatment type for any effect size included in the analysis (i.e., PND, PZD, and MBLR) was found. However, our findings are consistent with the results of several other meta-analyses in this domain (e.g., Heyvaert et al., 2012; Vanderkerken et al., 2013).

Another variable that is often found to be a statistically significant moderator in meta-analyses published in this domain is pretreatment functional analysis (e.g., Campbell, 2003;

Didden et al., 2006; Harvey et al., 2009; Kahng et al., 2002; Scotti et al., 1991). Consistent with these meta-analyses, we found a statistically significant relation between the overall intervention effect and the presence of functional analysis for four out of the five effect sizes (i.e., PND, MBLR, PEM, and PAND): Interventions preceded by a functional analysis reduced problem behavior significantly more than interventions not preceded by a functional analysis. Presence of functional analysis proved to be related to publication year and intervention type. More recently published studies were more likely to report on a functional analysis conducted prior to the behavioral intervention. The interventions most likely to be preceded by a functional analysis were respectively noncontingent reinforcement, antecedent exercise, DRA, combinations of positive interventions, combinations of aversive and positive interventions, and DRO interventions.

In our study as well as in the study of Campbell (2003), no main effect for target behavior for any effect size was found. Accordingly, we can conclude that the behavioral treatments were equally effective regardless of the type of problem behavior that was targeted. This finding is consistent with several other meta-analyses published in this domain as well (e.g., Didden et al., 2006; Heyvaert et al., 2010, 2014; Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004), although some meta-analyses do report on statistically significant moderating effects of behavior type (e.g., Heyvaert et al., 2012; Vanderkerken et al., 2013).

We conducted hierarchical multiple regression analyses for participant, intervention, and experimental characteristics to answer the third research question. The group of participant characteristics included in our study (i.e., age, gender, criteria used for diagnosing autism, intellectual disability level, and level of verbal communication ability) did not significantly influence the overall efficacy of the behavioral interventions. On the contrary, the group of treatment characteristics included in our study (i.e., parental involvement,

functional analysis, presence of follow up data, and efforts to generalize behavior change) did significantly influence the overall efficacy of the behavioral interventions. However, this was only true for the effect sizes PND, MBLR, PEM, and PAND, and not for PZD. The third group of characteristics, the experimental characteristics (i.e., design, number of coded baseline data points, number of coded treatment data points, publication year, and inter-rater reliability), was also found to significantly influence the overall efficacy of the behavioral interventions. Analogous to the results for the treatment characteristics, this was only true for the effect sizes PND, MBLR, PEM, and PAND, and not for PZD.

For this third question, our results for the participant variables with regard to the single-subject studies published between 1999 and 2012 correspond to the results of Campbell (2003) for the studies published between 1966 and 1998: The set of participant variables did not significantly account for additional variance in any effect size. In contrast with our statistically significant results for the treatment variables for the effect sizes PND, MBLR, PEM, and PAND, Campbell (2003) found that the treatment variables did not significantly account for additional variance in the effect sizes PND, PZD, and MBLR. With regard to the third group, we found that experimental characteristics significantly influenced the overall efficacy of the behavioral interventions for the effect sizes PND, MBLR, PEM, and PAND, and not for PZD. However, Campbell (2003) did find a statistically significant effect for PZD: The group of experimental variables accounted for 13% of the variance in PZD scores.

The fourth research question concerned the comparison of the effect sizes: We wanted to examine whether there were differences in the conclusions relating to treatment efficacy and moderating variables for the five calculated effect sizes. The analyses for PND, MBLR, PEM, and PAND resulted in similar conclusions related to treatment efficacy and moderating variables, whereas PZD resulted in contrary conclusions. There are several conceptual

differences that might contribute to the contrary findings for PZD on the one hand, and PND, MBLR, PEM, and PAND on the other hand. First of all, PND, MBLR, PEM, and PAND measure *reduction* of problem behavior, whereas PZD measures *complete suppression* of problem behavior (cf. **2.3**). Second, PZD only takes into account the data points in the intervention phase, whereas PND, MBLR, PEM, and PAND take into account data points in the baseline as well as the intervention phase. Third, PZD only takes into account the data points in the intervention phase from the first data point that reaches zero onwards, and discards all previous data points. In case there is no data point in the intervention phase that equals zero, PZD is zero. PND, MBLR, PEM, and PAND take into account all intervention phase data points.

Relating to the final research question, we noted an increase in the use of positive interventions over time, especially in the use of combinations of positive interventions, antecedent control interventions, and social stories interventions. Related to the increase in the use of positive interventions, we saw a decrease in the use of negative interventions over time, such as punishment and overcorrection interventions. Furthermore, although we a priori hypothesized it would be possible to find larger intervention effects in the more recently published set of studies due to higher-quality interventions, we found similar overall intervention effects reported for the present data set and the set of Campbell (2003). On the other hand, it has been argued that many scientifically discovered effects published in the literature seem to diminish with time (Schooler, 2011). However, we also did not see an overall decrease in the average effect sizes: The efficacy of behavioral interventions for reducing problem behavior in persons with autism did not seem to diminish over time.

The two most important implications for clinical practice of our study concern the questions which interventions could best (not) be used and promoted for reducing problem behavior in persons with autism, and what practitioners can do to increase the efficacy of behavioral interventions for reducing the problem behavior. Relating to the first issue, we found that positive combination interventions as well as aversive and positive combination interventions were statistically significantly better in reducing problem behavior than antecedent control only interventions. Furthermore, we found statistically significant detrimental evidence for the use of PECS only interventions. Relating to the second issue, we conclude that practitioners should continue to use behavioral interventions for reducing problem behavior in individuals with autism, because these interventions proved to be effective. In addition, we found that the interventions were equally effective regardless of the type of problem behavior that was targeted. Because interventions preceded by a functional analysis were found to reduce problem behavior significantly more than interventions not preceded by such an analysis, we advise practitioners to conduct pre-treatment functional analyses in order to increase the efficacy of the behavioral interventions.

Based on our findings, we want to formulate some methodological recommendations to substantial researchers and practitioners aiming to conduct single-subject studies in this field in the future. For respectively 62.8% and 79.3% of the 358 included participants, no attempts to generalize the behavior change and no follow-up data were reported. These high percentages are comparable to the ones reported in other meta-analyses of single-subject studies in this domain (e.g., Campbell, 2003; Didden et al., 2006; Heyvaert et al., 2012). However, the collection of data on the generalization of behavior changes as well as follow-up data are of utmost importance to study and document the durability of the changes across time and settings, and to demonstrate the functional utility of a treatment in extending beyond

the target behaviors or treatment environment into other areas of the participant's life (Tate et al., 2008). Accordingly, we strongly recommend single-subject researchers and practitioners to attempt to generalize the behavior change and to collect follow-up data. Furthermore, we are glad to see that the AB design was only used for 16 out of the 358 participants, and that most researchers used reversal and multiple baseline designs to study the efficacy of behavioral interventions for reducing problem behavior in persons with autism. According to the single-subject design standards developed by the What Works Clearinghouse, a single-subject study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions (Kratochwill et al., 2010). This implies that designs such as ABAB designs, multiple baseline designs with at least three baseline conditions, and alternating treatment designs with either at least three alternating treatments compared with a baseline condition or two alternating treatments compared with each other are considered to be credible single-subject designs, whereas AB, ABA, and BAB designs do not meet the design standards set forth by Kratochwill et al. (2010). Finally, it was reassuring to see that the mean number of observations in the first baseline and final treatment phases was 11 and 20, respectively. According to most single-subject methodologists, each baseline and treatment phase should consist of at least five repeated measures of the dependent variable in order to establish a pattern of response that can be used to predict future performance (i.e., for the baseline) and to convincingly document the effect of an intervention (Kratochwill et al., 2010; Wendt & Miller, 2012).

Like every study, our study has certain limitations. A first limitation concerns the fact that we did not have the resources to calculate the two more recently developed effect sizes (i.e., PEM and PAND) for the studies published between 1966 and 1998. Accordingly, in our answer to the final research question we could only compare the results for the effect sizes

PND, PZD, and MBLR. Another consideration relates to the question whether the average intervention effect is an adequate descriptive measure to summarize a distribution of effects. Alternative measures of central tendency are for instance the median and mode. Furthermore, it might also be interesting to study the variability, the skewness, and the kurtosis of the data.

Summarizing our study's results, we note eight key points. First of all, behavioral interventions were on average effective in reducing problem behavior in individuals with autism. Second, antecedent control only and PECS only interventions were less effective in reducing problem behavior, in comparison to the other interventions included in our study. Third, the use of positive interventions for reducing problem behavior in individuals with autism is increasing over time, whereas the use of negative interventions is decreasing. Fourth, the behavioral interventions were equally effective regardless of the type of problem behavior that was targeted. Fifth, interventions preceded by a functional analysis reduced problem behavior significantly more than interventions not preceded by a functional analysis. Accordingly, the observed trend that more recently published studies were more likely to include a functional analysis conducted prior to the behavioral intervention is definitely a positive one. Sixth, treatment and experimental characteristics, but not participant characteristics, were statistically significant moderators of the behavioral treatment effectiveness. Seventh, we found similar overall intervention effects reported for the single-subject studies published between 1966 and 1998 as for the ones published between 1999 and 2012. Finally, a methodological key point is that the PZD statistic resulted in conclusions contrary to the conclusions for the PND, MBLR, PEM, and PAND statistics.

References

Biosoft (1997-2014). *UnGraph Version 5* [Computer software]. Retrieved from http://www.biosoft.com/w/ungraph.htm

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behaviour in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24,* 120–138. doi:10.1016/S0891-4222(03)00014-3

Didden, R., Korzilius, H., van Oorsouw, W., & Sturmey, P. (2006). Behavioural treatment of challenging behaviours in individuals with mild mental retardation: Meta-analysis of single-subject research. *American Journal on Mental Retardation, 111,* 290–298.

Harvey, S. T., Boer, D., Meyer, L. H., & Evans, I. M. (2009). Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice. *Journal of Intellectual & Developmental Disability, 34,* 67–80. doi:10.1080/13668250802690922

Heyvaert, M., Maes, B., & Onghena, P. (2010). A meta-analysis of intervention effects on challenging behaviour among persons with intellectual disabilities. *Journal of Intellectual Disability Research, 54,* 634–649. doi:10.1111/j.1365-2788.2010.01291.x

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33,* 766–780. doi:10.1016/j.ridd.2011.10.010

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014). Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Focus on effectiveness in single-case experiments. *Journal of Applied Research in Intellectual Disabilities*. doi:10.1111/jar.12094

Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (in press). Randomization and data-analysis items in quality standards for single-case experimental studies. *Journal of Special Education.* doi:10.1177/0022466914525239

Kahng, S., Iwata, B. A., & Lewin, A. B. (2002). The impact of functional analysis on the treatment of self-injurious behavior. In S. R. Schroeder, M. L. Oster-Granite, & T. Thompson (Eds.), *Self-injurious behavior: Gene-brain-behavior relationships* (pp. 119–131). Washington, DC: American Psychological Association.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30,* 598–617. doi:10.1177/0145445504272974

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19,* 109–135. doi:10.1080/09362835.2011.565725

Matson, J. L., & LoVullo, S. V. (2008). A review of behavioral treatments for self-injurious behaviors of persons with autism spectrum disorders. *Behavior Modification, 32,* 61–76. doi:10.1177/0145445507304581

Matson, J. L., & LoVullo, S. V. (2009). Trends and topics in autism spectrum disorders research. *Research in Autism Spectrum Disorders, 3,* 252–257. doi:10.1016/j.rasd.2008.06.005

Murphy, O., Healy, O., & Leader, G. (2009). Risk factors for challenging behaviors among 157 children with autism spectrum disorder in Ireland. *Research in Autism Spectrum Disorders, 3,* 474–482. doi:10.1016/j.rasd.2008.09.008

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34,* 189–211. doi:10.1016/S0005-7894(03)80013-8

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40,* 194–204. doi:10.1177/00224669070400040101

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303–322. doi:10.1177/0145445511399147

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437. doi:10.1038/470437a

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96,* 233–256.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. Methodology and validation. *Remedial and Special Education, 8,* 24–33. doi:10.1177/074193258700800206

Shadish, W. R., Brasil, I. C. C., llingworth, D. A. I., White, K. D., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods, 41,* 177–183. doi:10.3758/BRM.41.1.177

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions, 6,* 228–237. doi:10.1177/10983007040060040401

SPSS Inc. (2013-2014). *SPSS for Windows Version 22* [Computer software]. Chicago, IL: SPSS Inc.

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation, 18,* 385–401. doi:10.1080/09602010802009201

Vanderkerken, L., Heyvaert, M., Maes, B., & Onghena, P. (2013). Psychosocial interventions for reducing vocal challenging behaviour in persons with autistic disorder: A multilevel meta-analysis of single-case experiments. *Research in Developmental Disabilities, 34,* 4515–4533. doi:10.1016/j.ridd.2013.09.030

Walsh, C. E., Mulder, E., & Tudor, M. E. (2013). Predictors of parent stress in a sample of children with ASD: Pain, problem behavior, and parental coping. *Research in Autism Spectrum Disorders, 7,* 256–264. doi:10.1016/j.rasd.2012.08.010

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35,* 235–268. doi:10.1353/etc.2012.0010

Table 1

*Participant Characteristics (N = 358)*

| Characteristic | *n* | Percent |
|---|---|---|
| **Age** (in years) (*M* = 10.24; *SD* = 8.06) | 353 | |
| **Gender** | | |
| Male | 286 | 79.9 |
| Female | 67 | 18.7 |
| Not reported | 5 | 1.4 |
| **Intellectual disability level** (IQ range) | | |
| None (IQ > 70) | 20 | 5.6 |
| Mild (70 – 55) | 14 | 3.9 |
| Moderate (54 – 40) | 30 | 8.4 |
| Severe / Profound (IQ < 40) | 48 | 13.4 |
| Not reported / Unclear | 246 | 68.7 |
| **Level of verbal ability** | | |
| Average language skills | 27 | 7.5 |
| Minimally verbal; some functional language | 115 | 32.1 |
| Nonverbal; mute | 61 | 17.0 |
| Not reported / Unclear | 155 | 43.3 |
| **Diagnostic criteria used** | | |
| DSM-III or DSM-III-TR | 1 | 0.3 |
| DSM-IV or DSM-IV-TR | 30 | 8.4 |
| ICD-10 | 3 | 0.8 |
| Not reported | 324 | 90.5 |

Table 2

*Treatment Characteristics (N = 358)*

| Characteristic | *n* | Percent |
|---|---|---|
| **Behavior targeted for reduction** (behavior grouping) | | |
| Internal and external behaviors combined | 78 | 21.8 |
| External combined | 37 | 10.3 |
| Internal combined | 1 | 0.3 |
| Stereotyped behavior only (internal) | 100 | 27.9 |
| Self-injurious behavior only (internal) | 33 | 9.2 |
| Disruptive behavior only (external) | 82 | 22.9 |
| Aggression only (external) | 26 | 7.3 |
| Property destruction only (external) | 1 | 0.3 |
| **Type of intervention** (treatment grouping) | | |
| Aversive and positive combinations | 74 | 20.7 |
| Positive combinations | 114 | 31.8 |
| Punishment only (aversive/punishment) | 10 | 2.8 |
| Differential reinforcement of other behavior only (positive) | 14 | 3.9 |
| Antecedent control only (positive) | 61 | 17.0 |
| Differential reinforcement of incompatible behavior only (positive) | 5 | 1.4 |
| Differential reinforcement of alternative behavior only (positive) | 23 | 6.4 |
| Antecedent exercise only (positive) | 2 | 0.6 |
| Noncontingent reinforcement only (positive) | 14 | 3.9 |
| Escape only (positive) | 1 | 0.3 |
| Social stories only (positive) | 29 | 8.1 |
| Picture exchange communication system (PECS) only (positive) | 5 | 1.4 |
| Mindfulness-based strategy only (positive) | 6 | 1.7 |
| **Functional analysis conducted** | | |
| Yes | 257 | 71.8 |
| No / Not reported | 101 | 28.2 |
| **Attempt to generalize behavior change** | | |
| Yes | 133 | 37.2 |
| No / Not reported | 225 | 62.8 |
| **Follow-up data collected** | | |
| Yes | 74 | 20.7 |
| No / Not reported | 284 | 79.3 |
| **Parent involved in treatment** | | |
| Yes | 56 | 15.6 |
| No / Not reported | 302 | 84.4 |

Table 3

*Experimental Characteristics (N = 358)*

| Characteristic | *n* | Percent |
|---|---|---|
| **Experimental design** | | |
| Reversal only | 128 | 35.8 |
| Multiple baseline only | 112 | 31.3 |
| AB only | 16 | 4.5 |
| Multiple baseline and Reversal | 8 | 2.2 |
| Alternating treatments only | 33 | 9.2 |
| Multiple baseline and Alternating treatments | 7 | 2.0 |
| Alternating treatments and Reversal | 34 | 9.5 |
| AB and Alternating treatments | 19 | 5.3 |
| Multiple baseline and Reversal and Alternating treatments | 1 | 0.3 |
| **Number of observations in first baseline phase** (*M* = 11.06; *SD* = 11.98) | | |
| **Number of observations in final treatment phase** (*M* = 19.92; *SD* = 22.85) | | |
| **Publication year** | | |
| 1999 | 10 | 2.8 |
| 2000 | 19 | 5.3 |
| 2001 | 13 | 3.6 |
| 2002 | 23 | 6.4 |
| 2003 | 7 | 2.0 |
| 2004 | 17 | 4.7 |
| 2005 | 25 | 7.0 |
| 2006 | 8 | 2.2 |
| 2007 | 32 | 8.9 |
| 2008 | 22 | 6.1 |
| 2009 | 46 | 12.8 |
| 2010 | 28 | 7.8 |
| 2011 | 70 | 19.6 |
| 2012 | 38 | 10.6 |
| **Inter-rater reliability data** | | |
| Yes | 333 | 93.0 |
| No / Not reported | 25 | 7.0 |

Table 4

*Effect Sizes Calculated at the Participant and the Study Level*

| Effect size | Participant level | Study level |
|---|---|---|
| PND | *M* = .75; *SD* = .34<br>95% CI [.71, .78] | *M* = .76; *SD* = .30<br>95% CI [.72, .80] |
| PZD | *M* = .45; *SD* = .36<br>95% CI [.41, .48] | *M* = .47; *SD* = .35<br>95% CI [.43, .52] |
| MBLR | *M* = .70; *SD* = .40<br>95% CI [.66, .74] | *M* = .74; *SD* = .30<br>95% CI [.70, .78] |
| PEM | *M* = .91; *SD* = .19<br>95% CI [.89, .93] | *M* = .93; *SD* = .14<br>95% CI [.91, .95] |
| PAND | *M* = .92; *SD* = .11<br>95% CI [.91, .93] | *M* = .92; *SD* = .10<br>95% CI [.91, .94] |

*Note*. PND = percentage of nonoverlapping data; PZD = percentage of zero data; MBLR = mean baseline reduction; PEM = percentage of data points exceeding the median; PAND = percentage of all nonoverlapping data; CI = confidence interval.

Table 5

*Two-way ANOVA Results for Effect Sizes*

| Effect size | Source | *df1* | *df2* | *F* | *p* |
|---|---|---|---|---|---|
| PND | Treatment type | 12 | 307 | 3.95 | < .001* |
| | Behavior category | 7 | 307 | 1.28 | .262 |
| | Treatment by behavior | 31 | 307 | 1.50 | .047 |
| | Total model | 50 | 307 | 2.09 | < .001* |
| PZD | Treatment type | 12 | 307 | 2.03 | .021 |
| | Behavior category | 7 | 307 | 1.54 | .152 |
| | Treatment by behavior | 31 | 307 | 0.87 | .668 |
| | Total model | 50 | 307 | 2.07 | < .001* |
| MBLR | Treatment type | 12 | 307 | 2.05 | .020 |
| | Behavior category | 7 | 307 | 0.91 | .503 |
| | Treatment by behavior | 31 | 307 | 0.96 | .527 |
| | Total model | 50 | 307 | 1.62 | .008 |
| PEM | Treatment type | 12 | 307 | 3.32 | < .001* |
| | Behavior category | 7 | 307 | 1.18 | .313 |
| | Treatment by behavior | 31 | 307 | 0.89 | .641 |
| | Total model | 50 | 307 | 2.00 | < .001* |
| PAND | Treatment type | 12 | 307 | 3.64 | < .001* |
| | Behavior category | 7 | 307 | 0.80 | .592 |
| | Treatment by behavior | 31 | 307 | 1.14 | .284 |
| | Total model | 50 | 307 | 2.18 | < .001* |

*Note.* PND = percentage of nonoverlapping data; PZD = percentage of zero data; MBLR = mean baseline reduction; PEM = percentage of data points exceeding the median; PAND = percentage of all nonoverlapping data.

* = Statistically significant *p* value after Bonferroni correction for multiple testing: Results associated with *p* values below .003 (.05/15) were interpreted as statistically significant.

Table 6

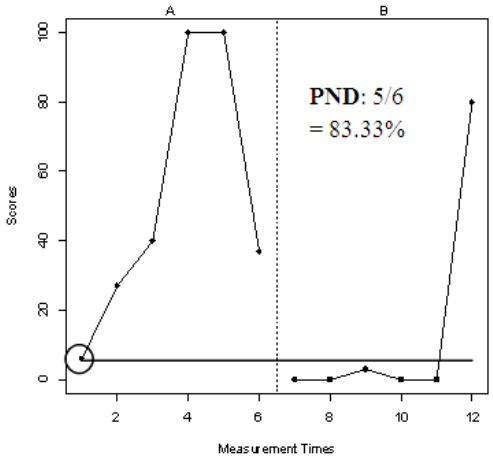
*Results of the Hierarchical Multiple Regression Analyses*

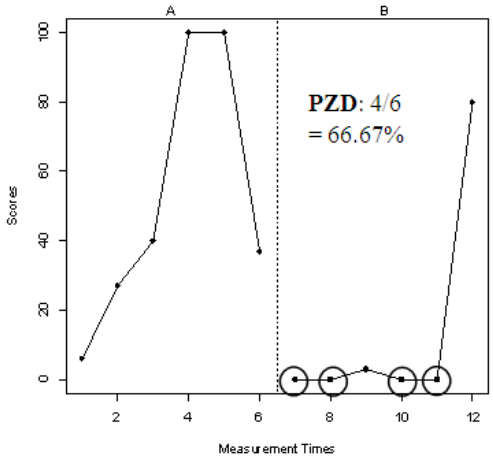
| Predictors included | Effect size | Model summary |
|---|---|---|
| 1. Participant characteristics<br>2. Treatment type and Behavior category | PND | Model 1: Predictors A B C D E F G; *F Change*(7, 345) = 2.22, *p* = .032<br>Model 2: Predictors A B D E F G; *F Change*(1, 345) = 0.04, *p* = .846<br>Model 3: Predictors A B D F G; *F Change*(1, 346) = 0.41, *p* = .522<br>Model 4: Predictors A B D G; *F Change*(1, 347) = 0.66, *p* = .418<br>Model 5: Predictors A B G; *F Change*(1, 348) = 0.85, *p* = .358<br>Model 6: Predictors A B; *F Change*(1, 349) = 1.18, *p* = .278 |
| | PZD | Model 1: Predictors A B C D E F G; *F Change*(7, 345) = 0.86, *p* = .540<br>Model 2: Predictors A B C D E F; *F Change*(1, 345) = 0.00, *p* = .963<br>Model 3: Predictors A B C D E; *F Change*(1, 346) = 0.02, *p* = .900<br>Model 4: Predictors A B C E; *F Change*(1, 347) = 0.78, *p* = .378<br>Model 5: Predictors A B E; *F Change*(1, 348) = 1.30, *p* = .255 |
| | MBLR | Model 1: Predictors A B C D E F G; *F Change*(7, 345) = 1.64, *p* = .122<br>Model 2: Predictors A B C E F G; *F Change*(1, 345) = 0.07, *p* = .793<br>Model 3: Predictors A B C E F; *F Change*(1, 346) = 0.41, *p* = .523<br>Model 4: Predictors A B C E; *F Change*(1, 347) = 0.65, *p* = .422<br>Model 5: Predictors A B C; *F Change*(1, 348) = 1.72, *p* = .190<br>Model 6: Predictors A B; *F Change*(1, 349) = 2.27, *p* = .132 |
| | PEM | Model 1: Predictors A B C D E F G; *F Change*(7, 345) = 1.69, *p* = .110<br>Model 2: Predictors A B C D E G; *F Change*(1, 345) = 0.01, *p* = .911<br>Model 3: Predictors A B C E G; *F Change*(1, 346) = 0.01, *p* = .917<br>Model 4: Predictors A B C E; *F Change*(1, 347) = 0.05, *p* = .831<br>Model 5: Predictors A B C; *F Change*(1, 348) = 0.23, *p* = .630<br>Model 6: Predictors A B; *F Change*(1, 349) = 0.97, *p* = .327 |
| | PAND | Model 1: Predictors A B C D E F G; *F Change*(7, 345) = 2.10, *p* = .043<br>Model 2: Predictors A B C D E G; *F Change*(1, 345) = 0.02, *p* = .885<br>Model 3: Predictors A B C E G; *F Change*(1, 346) = 0.04, *p* = .843<br>Model 4: Predictors A B E G; *F Change*(1, 347) = 0.07, *p* = .796<br>Model 5: Predictors A B E; *F Change*(1, 348) = 0.10, *p* = .757<br>Model 6: Predictors A B; *F Change*(1, 349) = 0.18, *p* = .676 |
| 1. Treatment characteristics<br>2. Treatment type and Behavior category | PND | Model 1: Predictors A B H I J K; *F Change*(6, 351) = 6.69, *p* < .001*<br>Model 2: Predictors A B H I K; *F Change*(1, 351) = 0.01, *p* = .915 |
| | PZD | Model 1: Predictors A B H I J K; *F Change*(6, 351) = 2.60, *p* = .018<br>Model 2: Predictors A B H J K; *F Change*(1, 351) = 0.01, *p* = .919<br>Model 3: Predictors A B H J; *F Change*(1, 352) = 0.45, *p* = .502 |
| | MBLR | Model 1: Predictors A B H I J K; *F Change*(6, 351) = 3.89, *p* = .001*<br>Model 2: Predictors A B I J K; *F Change*(1, 351) = 0.98, *p* = .322<br>Model 3: Predictors A B J K; *F Change*(1, 352) = 0.95, *p* = .331 |
| | PEM | Model 1: Predictors A B H I J K; *F Change*(6, 351) = 6.34, *p* < .001*<br>Model 2: Predictors A B H I K; *F Change*(1, 351) = 0.79, *p* = .376<br>Model 3: Predictors A B I K; *F Change*(1, 352) = 1.30, *p* = .255 |

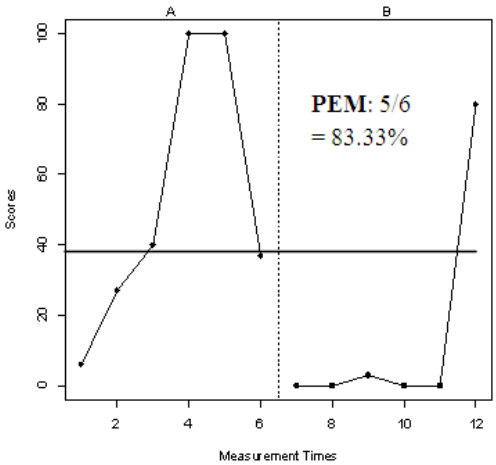
| | | |
|---|---|---|
| | PAND | Model 1: Predictors A B H I J K; *F Change*(6, 351) = 5.52, $p < .001$*<br>Model 2: Predictors A B H I K; *F Change*(1, 351) = 0.01, $p = .909$<br>Model 3: Predictors A B I K; *F Change*(1, 352) = 1.34, $p = .249$ |
| 1. Experimental characteristics<br>2. Treatment type and Behavior category | PND | Model 1: Predictors A B L M N O P; *F Change*(7, 350) = 12.90, $p < .001$*<br>Model 2: Predictors A B L M N P; *F Change*(1, 350) = 0.00, $p = .987$<br>Model 3: Predictors A B L M P; *F Change*(1, 351) = 1.21, $p = .273$ |
| | PZD | Model 1: Predictors A B L M N O P; *F Change*(7, 350) = 2.15, $p = .038$<br>Model 2: Predictors A B L M N O; *F Change*(1, 350) = 0.23, $p = .630$<br>Model 3: Predictors A B L M O; *F Change*(1, 351) = 0.10, $p = .756$<br>Model 4: Predictors A B L M; *F Change*(1, 352) = 1.19, $p = .275$ |
| | MBLR | Model 1: Predictors A B L M N O P; *F Change*(7, 350) = 4.22, $p < .001$*<br>Model 2: Predictors A B L N O P; *F Change*(1, 350) = 0.35, $p = .553$<br>Model 3: Predictors A B L N P; *F Change*(1, 351) = 1.71, $p = .192$ |
| | PEM | Model 1: Predictors A B L M N O P; *F Change*(7, 350) = 5.72, $p < .001$*<br>Model 2: Predictors A B L M N P; *F Change*(1, 350) = 0.17, $p = .678$<br>Model 3: Predictors A B L M P; *F Change*(1, 351) = 2.57, $p = .110$ |
| | PAND | Model 1: Predictors A B L M N O P; *F Change*(7, 350) = 10.22, $p < .001$*<br>Model 2: Predictors A B L M N P; *F Change*(1, 350) = 0.29, $p = .593$<br>Model 3: Predictors A B M N P; *F Change*(1, 351) = 1.76, $p = .185$ |

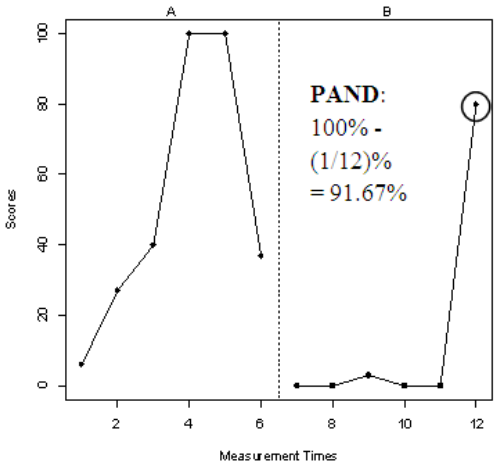
*Note.* A = treatment type; B = behavior category; C = criteria used for diagnosing autism; D = gender; E = intellectual disability level; F = level of verbal communication ability; G = age; H = efforts to generalize behavior change; I = presence of follow up data; J = parental involvement in treatment; K = functional analysis; L = publication year; M = inter-rater reliability; N = number of coded treatment data points; O = design; P = number of coded baseline data points; PND = percentage of nonoverlapping data; PZD = percentage of zero data; MBLR = mean baseline reduction; PEM = percentage of data points exceeding the median; PAND = percentage of all nonoverlapping data.

* = Statistically significant *p* value after Bonferroni correction for multiple testing: Results associated with *p* values below .003 (.05/15) were interpreted as statistically significant.




| Phase | Dependent variable |
|---|---|
| A | 6 |
| A | 27 |
| A | 40 |
| A | 100 |
| A | 100 |
| A | 37 |
| B | 0 |
| B | 0 |
| B | 3 |
| B | 0 |
| B | 0 |
| B | 80 |

Mean baseline value: 51.67

Mean treatment value: 13.83

**MBLR**: (51.67 - 13.83) / 51.67

= 73.23%

*Figure 1*. Calculation of the five effect sizes: An empirical example.